

AD-A079 000

DEFENSE TECHNICAL INFORMATION CENTER ALEXANDRIA VA  
RETRIEVAL PERFORMANCE OF A KWOC INDEX TO THE 1498 WORK UNIT SUM--ETC(U)  
JAN 79 E P BURRESS  
DTIC/TR-79/2

F/G 5/2

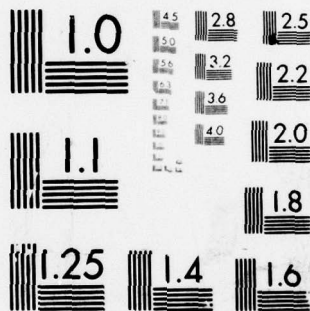
UNCLASSIFIED

NL

| OF |  
AD  
A079000



END  
DATE  
FILMED  
2-80  
DDC



MICROCOPY RESOLUTION TEST CHART  
NATIONAL BUREAU OF STANDARDS-1963-A

DTIC/TR-79/2

LEVEL II (1)

AD-A079 000

AD-A079000

**RETRIEVAL PERFORMANCE OF A KWOC  
INDEX TO THE 1498 WORK UNIT SUMMARY  
DATA BASE**

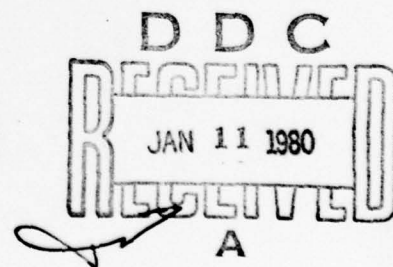
DDC FILE COPY

OFFICE OF SUPPORT SERVICES

JANUARY 1979

Approved for public release; distribution unlimited

**DEFENSE TECHNICAL INFORMATION CENTER  
Defense Logistics Agency  
Cameron Station  
Alexandria, VA 22314**



| REPORT DOCUMENTATION PAGE  |                                      | READ INSTRUCTIONS<br>BEFORE COMPLETING FORM                              |
|--|--------------------------------------|--|
| 1. REPORT NUMBER<br><b>14</b> DTIC/TR-79/2   | 2. GOVT ACCESSION NO.<br>AD-A079 000 | 3. RECIPIENT'S CATALOG NUMBER  |
| 4. TITLE (and Subtitle)<br>Retrieval Performance of a KWOC Index to the<br>1498 Work Unit Summary Data Base  |                                      | 5. TYPE OF REPORT & PERIOD COVERED                                       |
| 7. AUTHOR(s)<br><b>10</b> Elaine Pontani/Burress   |                                      | 6. PERFORMING ORG. REPORT NUMBER   |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Defense Technical Information Center<br>Cameron Station<br>Alexandria, Virginia 22314   |                                      | 10. PROGRAM ELEMENT, PROJECT, TASK<br>AREA & WORK UNIT NUMBERS<br>65801S |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br><b>12/58</b>  |                                      | 12. REPORT DATE<br><b>11</b> January 1979                                |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)<br><b>11 Jan 80</b>  |                                      | 13. NUMBER OF PAGES<br>27  |
|  |                                      | 15. SECURITY CLASS. (of this report)<br>Unclassified                     |
|  |                                      | 15a. DECLASSIFICATION/DOWNGRADING<br>SCHEDULE                            |
| 16. DISTRIBUTION STATEMENT (of this Report)<br>Approved for public release; distribution unlimited.  |                                      |  |
| 17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)   |                                      |  |
| 18. SUPPLEMENTARY NOTES  |                                      |  |
| 19. KEY WORDS (Continue on reverse side if necessary and identify by block number)<br>Information Retrieval<br>Machine-Aided Indexing<br>KWOC (Key-Word Out of Context)  |                                      |  |
| 20. ABSTRACT (Continue on reverse side if necessary and identify by block number)<br>The retrieval performance of a technique for automated indexing using a controlled vocabulary (machine-aided indexing) is compared with that of a Key Word Out of Context index (KWOC). Twenty subject searches were performed using a data base of 635 summaries of Defense management data. Results show an average recall measurement of .73 for machine-aided indexing and .79 for the KWOC index. The average relevance measurements were .62 for machine-aided indexing and .60 for the KWOC index. |                                      |  |



**SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)**

**SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)**



**DEFENSE LOGISTICS AGENCY**  
DEFENSE TECHNICAL INFORMATION CENTER  
CAMERON STATION  
ALEXANDRIA, VIRGINIA 22314

DTIC-B

Jan 79

**PREFACE**

The Defense Technical Information Center (DTIC) has long been interested in indexing concepts as one aspect of its information processing activities. Over ten years ago DTIC began investigating ways that automation could be used in indexing. One outgrowth of this interest is Machine-Aided Indexing (MAI) which DTIC has used to index three of its data bases. This paper compares the retrieval performance of Machine-Aided Indexing and a Key Word Out of Context index (KWOC). This is not meant to be a definitive study but rather to be informative and possibly spur further research.

This report is a summary of research conducted by a member of DTIC's Information Sciences Intern Program (ISIP). The ISIP is a two-year training program which consists of rotational assignments throughout the agency and requires the performance of research studies, usually pertaining to one aspect of DTIC's operations. Even though this research was initiated to meet specific program requirements, it is felt that it will also be of interest to the DTIC user community.

Prepared by:

*Elaine P. Burress*

ELAINE P. BURRESS  
Librarian

Approved by:

*Betty H. Hughes*

*for* RENE' S. LEHMAN  
Director, Office of  
Support Services

|                    |  |
|--------------------|--|
| Accession For      |  |
| NTIS GRA&I         | <input checked="checked" type="checkbox"/> |
| DDC TAB            | <input type="checkbox"/>                   |
| Unannounced        | <input type="checkbox"/>                   |
| Justification      |  |
| By                 |  |
| Distribution/      |  |
| Availability Codes |  |
| Dist               | Avail and/or special                       |
| A                  |  |

## INDEX

|   |    |
|---|----|
| Introduction.....                                 | 1  |
| The KWOC Index.....                               | 3  |
| Procedure.....                                    | 6  |
| Recall and Relevance.....                         | 9  |
| Results.....                                      | 13 |
| Exhibits  |    |
| A: Test Recall and Relevance Results Summary..... | 16 |
| B: Recall Failures.....                           | 17 |
| C: Relevance Failures.....                        | 18 |
| D: Recall and Relevance Failures Per Search ..... | 19 |
| Discussion.....                                   | 20 |
| Selected Bibliography.....                        | 22 |

## INTRODUCTION

This work investigates the impact on subject retrieval in the 1498 Work Unit Summary Data Base by incorporating the ability to search for documents using terms from the title. The specific question that this study attempts to answer is: How do index terms taken directly from the title affect the retriever's ability to retrieve documents?

Interest in the use of index terms taken directly from document titles was sparked by the realization that the three commercial data base vendors, Lockheed, Systems Development Corporation and Bibliographic Retrieval Service, who provide access to our collection of unclassified, unlimited technical report citations via National Technical Information Service (NTIS), provide "full text" searching of the title, and in certain cases, abstracts. Full text searching is provided by allowing the significant words of the text, in this case - titles and possibly abstracts, to be used as index terms. Essentially, a KWOC (Key Word Out of Context) index is created and available to be searched on-line. The commercial data base vendors also provide for searching using manually assigned index terms. Our unclassified, unlimited collection of technical reports can be searched using descriptors and identifiers assigned at DTIC and enhanced by NTIS and the key words extracted from titles, and in some cases, abstracts.



Two questions concerning the use of full text search capabilities are of interest to DTIC - does the full text search capability, combined with traditional indexing, improve retrieval and how does full text searching of the title compare with searching using terms assigned by Machine-Aided Indexing (MAI)? Answers to this second question will be pursued in this paper.

## THE KWOC INDEX

The concept of preparing an index to a document using significant words from the title has been around about 25 years.<sup>1</sup> Hans Peter Luhn of IBM, working in the mid-1950's, used data processing equipment to generate the indexing and promoted two kinds of indexing formats. The Key Word Out of Context (KWOC) format, the one used for this report, displays the index terms along with information identifying the document to which it refers.

There are many enthusiastic supporters of this kind of indexing. The major reason for its acceptance is the speed at which the index is produced. This speed is the result of the elimination of human intellectual effort and the use of computer processing. In many cases, greater speed and timeliness are achieved at significantly lower cost. Another advantage which is typically claimed for KWOC indexes is the use of the author's own terminology. It is felt that the author, a member of the community with which he wants to communicate and intimately familiar with the material being indexed, is best able to describe his or her document.<sup>2</sup>

<sup>1</sup> Gerald Jahoda, Information Storage and Retrieval Systems for Individual Researchers (New York: Wiley-Interscience, 1970), p. 83.

<sup>2</sup> Mary E. Stevens, Automatic Indexing: A State-of-the-Art (Washington, D.C.: National Bureau of Standards, 1965), p. 55-67.



The most common type of complaint against the KWOC indexing method is the lack of terminological control. The familiar problems associated with searching using all synonyms, near-synonyms or variants under which a concept may be indexed, are intensified by lack of a thesaurus. In addition, the normal difficulty of matching searcher - indexer language is aggravated by the multiplicity of "indexers".<sup>1</sup> Another major concern is the adequacy of just the title to generate the indexing. By their nature, titles describe only the principle subject of the document. Consequently, a KWOC title index cannot provide access to minor subjects discussed in the document.

Despite these difficulties, several data bases are indexed using KWOC techniques and the major data base vendors offer this capability as an enhancement to traditional indexing. For example, BioSciences Information Service (BIOSIS) has been preparing permuted title-fragment indexes for titles published in Biological Abstracts since 1959, and BioResearch Index (BioRI) since 1967.<sup>2</sup>

The portion of our collection released to the public by NTIS can be accessed with words from the title through Lockheed's

<sup>1</sup>Stevens, p. 55-67.

<sup>2</sup>Maureen Lefever, "Managing an Uncontrolled Vocabulary Ex Post Facto," Journal of the American Society for Information Science, 23:6 (November-December, 1972), p. 339.

Dialog system. System Development Corporation's Orbit system permits searching of the NTIS data base using words from the title. In addition, once a subset of the collection has been designated as relevant, it can be further inspected using words from the abstract text. According to its promotional literature, the Bibliographic Retrieval Service program uses an enhanced version of STAIRS (IBM's Storage and Information Retrieval System). It permits searching of titles and abstracts for words, phrases or numbers and allows the searcher to specify the exact positional relationship of one term to another in a document by using logical operators.

According to Curt L. Harris and Suzanne H. Roberts of GE, "The ultimate answer to the user's need is the full-text search." To answer this need GE has developed hardware, the Associative Processor, designed to manipulate large amounts of text. The advantages claimed for full text searching are the elimination of the inverted file, the speed of preparation, and searchability of any text file, not just those designed for the system.<sup>1</sup>

<sup>1</sup>Curt L. Harris and Suzanne H. Roberts, "The Search for Tomorrow: Low Cost, Full Text Searching with Minimum Front End Investment," Proceedings of the ASIS Annual Meeting (White Plains, New York, 1978), p. 368.

## PROCEDURE

In this study, a KWOC title index was prepared for a sample of 635 Work Unit Summaries (1498's) which had previously been indexed with MAI. The KWOC index created for this report displays the index terms, the accession number of the document to which the term refers, and the document's full title. A COBOL computer program was written to select English words from titles in the 1498 data base. These words were matched against a stop-list of insignificant words which are not useful index terms. If the selected word is not on the stop-list, the program printed the selected word as an index term along with the accession number of the document and the full title. This process was repeated for all English words in the titles of the 1498's in the sample. For example, for the 1498 entitled "Computers in Information Sciences: Computer Components" the output would look like this:

|             |          |                                    |                     |
|-------------|----------|------------------------------------|---------------------|
| Components  | DN123456 | Computers in Information Sciences: | Computer Components |
| Computer    | DN123456 | Computers in Information Sciences: | Computer Components |
| Computers   | DN123456 | Computers in Information Sciences: | Computer Components |
| Information | DN123456 | Computers in Information Sciences: | Computer Components |
| Sciences    | DN123456 | Computers in Information Sciences: | Computer Components |

The word "in" is on the stop-list and would not appear as an index term.

The extracted index terms for the documents are displayed alphabetically creating a manual index.

Twenty subject searches were performed using the sample on the on-line terminal with a search strategy based on the Defense Retrieval and Indexing Terminology (DRIT) and the capabilities of masking, searching an index term's hierarchy and weighting index terms. Each of the 20 searches was repeated using a different search strategy and a

manual index of keywords (the KWOC) generated by a COBOL program written for this purpose. Search results were analyzed to uncover:

a. How do KWOC title indexing and MAI compare in their ability to allow retrieval of all documents?

b. How do they compare in their ability to withhold nonrevelant documents?

c. Does it appear that KWOC title indexing is equally, less, or more effective than MAI?

The 1498 Work Unit Summaries were chosen for this study for three reasons. Presently, DTIC does not have access to commercial data bases which provide free text search capabilities. There are plans to make these data bases available in the near future. Since some of the 1473 Technical Report Data Base is available on these commercial systems, the efficacy of free text search capabilities for the 1473 can be best explored when commercial systems are available rather than simulating free text searching in-house. Since commercial data base vendors do not provide access to the 1498 data base, simulating the free test search in-house will not duplicate work done elsewhere. Secondly, the 1498 data base is indexed with MAI and allows comparison of MAI technique with the more traditional KWOC. Lastly, since the complete 1498 record is stored on-line, the entire record is easily available for making relevance judgments.



The Work Unit Information System Data Bank (1498) is a collection of summaries of research and development performed at the work unit level by the Department of Defense and its contractors. The entire summary is stored in the computer; narrative fields are the title, technical objective, the approach and progress. The 1498 summaries are indexed by DTIC with MAI. The MAI program reads all of the narrative fields, recognizes both index terms and use references of the DTIC Natural Language Data Base (NLDB), and assigns the appropriate index terms to the summary. A listing of words and phrases not in the DTIC NLDB is also generated. DTIC policy is to manually review both index terms assigned, and words and phrases not found in the NLDB, and to make changes in the index term assignments as appropriate. This sample was not manually reviewed.

MAI is similar to KWOC indexing in that both use computers to read text to generate terms. The MAI is far more sophisticated than KWOC in that it uses a controlled vocabulary.<sup>1,2</sup>

Another difference between MAI and KWOC, in this instance is that MAI uses all narrative fields, whereas, the KWOC index is only a full text search of the title.

<sup>1</sup>Charles R. Jacobs, Machine-Aided Indexing: Technical Progress Report for Period July 1971-June 1972 (Alexandria, VA.: Defense Documentation Center, 1972).

<sup>2</sup>Paul H. Klingbiel, Machine-Aided Indexing: Technical Progress Report for Period Jan. 1967-June 1969 (Alexandria, VA.: Defense Documentation Center, 1969).

## RECALL AND RELEVANCE

Information retrieval system performance is normally evaluated in terms of its ability to retrieve material pertinent to a user's needs (relevance) and its ability to deliver everything it holds that is relevant to the request (recall).

Cleverdon in the introduction to the Cranfield Research Project writes:

"The reason why so much attention has been given to recall and relevance is that these are the only two user criteria which demand any serious effort in their measurement. They are concerned with whether the system is capable of locating what is sought. The unarguable fact is that they are fundamental requirements of the users, and it is quite unrealistic to try to measure how effectively a system is operating without bringing in recall and relevance".<sup>1</sup> Exactly how to measure recall and relevance and the reliability of these measurements have, however, been the subject of some controversy.

Relevance is expressed as the ratio of the number of documents retrieved which are considered relevant to the total number of documents retrieved. The ultimate test of a retrieval system, of course, is whether or not it satisfies the user. Consequently, the best judge of relevance is the user who posed the original search

<sup>1</sup>Cyril Cleverdon, Jack Mills and Michael Keen, ASLIB Cranfield Research Project: Factors Determining the Performance of Indexing Systems (Cranfield Bedfordshire: College of Aeronautics, 1966), p. 5.



question. In actual practice and in the ideal experiment, the user will pose the question and review the search results making relevance judgments.

In the present experiment, relevance was judged under less than ideal conditions. However, attempts were made to obtain accurate relevance figures using a technique similar to that described by Lancaster<sup>1</sup> using source documents. A source document is a document which would be considered 100% relevant to a user if it were retrieved in a subject search. A subject search is formulated based on the content of the source document. After the search is completed, the retrieved document's subject contents are compared to that of the source document and a relevance judgment is made based on the similarity between the two documents. For example, consider a subject search based upon the source document, "The Vegetarian Epicure". This subject search would specify those aspects of the source document that made its information relevant, say, that it contained recipes for vegetarian meals. Documents retrieved via the search containing recipes for vegetable meals would be relevant; documents containing recipes for vegetables as side dishes, general recipe books or books about growing vegetables would not be considered relevant.

<sup>1</sup>Frederick Lancaster, Information Retrieval Systems: Characteristics, Testing, and Evaluation (New York: John Wiley and Sons, Inc., 1965), p. 124-126.

In this experiment, 20 Work Unit Summaries (1498) were randomly selected as source documents. Both search strategies and relevance judgments were based on these source documents.

Recall is expressed as the ratio of the number of relevant documents retrieved by the system to the total number of relevant documents in the system. What we are looking for is the number of relevant documents not retrieved by the system and this is not easy to determine or even estimate. Theoretically, the only way to identify all of the documents in a system which would be relevant to a search query, would be to examine all of the documents. Even for a small data base this procedure is time consuming. Traditional ways to estimate the total number of relevant documents within a system are to search for these documents several times using different strategies. In this experiment, recall is estimated as the total number of relevant documents retrieved with MAI, plus the relevant documents retrieved by searching using the KWOC.

In the present study, each item retrieved was examined to see if its subject content was similar to that of the source document. Items were judged either relevant or nonrelevant. If the item was considered not relevant, reasons for its retrieval were sought by examining the search request, the index terms assigned to the retrieved document or its title. Reasons for failure were categorized as arising from faulty indexing vocabulary, incorrect indexing or inadequate searching procedures. Indexing vocabulary failures were further broken down into these categories:

lack of term specificity;  
term ambiguity;  
false coordination of terms; and  
failure to accommodate synonyms.

Indexing failures were recognized as:

failure to assign the appropriate term;  
assignment of inappropriate terms; and  
failure to assign terms at the appropriate level of exhaustivity.

Searching failures were seen as resulting from:

use of an inappropriate term;  
a too specific search strategy;  
a too general search strategy;  
failure to cover all approaches; and  
masking errors.

Recall failures were analyzed using the same criteria.

## RESULTS

As shown on the chart on page 16, the average recall measurements are .73 for MAI and .79 for the KWOC index. The relevance measurements are similarly close: .62 for MAI and .60 for the KWOC index. These recall and performance figures indicate that in this study, the MAI and KWOC index performed equally well. Both were equally able to retrieve all relevant documents and screen out nonrelevant documents.

An immediate observation is that the KWOC results show that the Work Unit Summary titles contain descriptive information and use words which make adequate index terms.

Analysis of the individual recall and relevance failures permits the following observations.

Recall failures: With KWOC indexing, recall failures occurred primarily because the KWOC failed to accommodate synonyms and the search strategies failed to cover all approaches. This simply means that a concept can be expressed in a title several ways, and since a KWOC provides no vocabulary guidance, for example, use references, unless the search expresses the concept in every possible way, recall will be less than ideal. For example, the subject of search #13 is radiation induced vascular damage to the internal capillaries of the eye. The terms "vascular", "damage", "capillaries", "eye", and "eyes" were searched as single terms in the KWOC index. The relevant document entitled "Acute ocular response to infection and radiation" was not retrieved since eye damage is expressed here as ocular response.



The primary cause of recall failures with MAI is the failure of MAI to post the proper term to the document. This did not occur very often, but it occurred most often (in 6 searches) and implies that some of the use references are faulty. These failures also reflect difficulties encountered constructing search strategy using the DRIT.

More searches using the KWOC index had perfect recall. Search #3 shows an interesting result; when 18 documents were deemed relevant there were 14 KWOC recall failures. This means that 14 relevant documents in the sample were not retrieved by the search using the KWOC index. This points out the difficulties which can be encountered when using a KWOC index to search a large number of relevant documents. The more relevant documents; the more complex the KWOC search must be in order to express a concept in several ways to match the different ways the concept is handled in the titles. In the remainder of the searches, in which it was ascertained that there were between 1 and 7 relevant documents in the sample used here, KWOC recall equalled, in fact was slightly better, than that using MAI.

Relevance failures: Searches using MAI retrieved nonrelevant documents primarily because both the assigned index terms and search strategies were too general. Added to this was a significant amount of ambiguity in index terms. Consequently, in the attempt to make the net big enough for all relevant documents to be caught, several nonrelevant documents were retrieved.

The subject of search #11 was survivability of equipment and personnel in nuclear warfare. The search was simply for all items indexed under Survival (General) or Combat Areas. A more specific search design - Survival (General) or Combat Areas and Nuclear Bombs or Nuclear Clouds or %Nuclear Explosions or %Nuclear W retrieved only one item. The simplified search retrieved 23 documents of which 7 were relevant. The bulk of the nonrelevant documents were indexed under Survival (General) and discussed either skin graft survival or spacecraft survival.

This work showed that a search with MAI can be highly specific. This is an unexpected result, as it was expected with using general index terms a search with MAI could not isolate one particular document. This proved not to be the case. In fact, very specific searches can be constructed which will retrieve only the source document.

In 2 of the 20 searches, searches with MAI had 15 and 16 relevance failures. Of the remaining searches, 13 had perfect relevance scores. MAI performed slightly better than the KWOC index in retrieving only relevant documents.

The primary reason for KWOC relevance failures was the failure of the KWOC index to provide adequate access to concepts described by synonymous phrases. Hand-in-hand with this was the failure of the search strategy to use the appropriate terms and specificity. Simply put, the searches failed to use the specific words and phrases the authors used in their titles to describe the document.



# TEST RECALL AND RELEVANCE RESULTS SUMMARY

| Search # | MAI    |           | KWOC   |           |
|----------|--------|-----------|--------|-----------|
|          | Recall | Relevance | Recall | Relevance |
| 1        | 1/1    | 1/1       | 1/1    | 1/1       |
| 2        | 1/2    | 1/1       | 1/2    | 1/3       |
| 3        | 17/18  | 17/33     | 4/18   | 4/4       |
| 4        | 1/1    | 1/1       | 1/1    | 1/1       |
| 5        | 3/4    | 3/3       | 4/4    | 4/4       |
| 6        | 1/1    | 1/1       | 1/1    | 1/1       |
| 7        | 0/1    | 0/2       | 0/1    | 0/0       |
| 8        | 2/2    | 2/2       | 2/2    | 2/2       |
| 9        | 4/4    | 4/5       | 3/4    | 3/4       |
| 10       | 1/1    | 1/5       | 1/1    | 1/2       |
| 11       | 7/7    | 7/23      | 5/7    | 5/11      |
| 12       | 0/3    | 0/2       | 3/3    | 3/8       |
| 13       | 0/1    | 0/0       | 0/1    | 0/2       |
| 14       | 0/1    | 0/0       | 1/1    | 1/2       |
| 15       | 1/1    | 1/4       | 1/1    | 1/5       |
| 16       | 2/3    | 2/2       | 2/3    | 2/2       |
| 17       | 1/1    | 1/1       | 1/1    | 1/1       |
| 18       | 1/1    | 1/1       | 1/1    | 1/3       |
| 19       | 1/1    | 1/1       | 1/1    | 1/2       |
| 20       | 1/1    | 1/2       | 1/1    | 1/4       |
| Average  | .73    | .62       | .79    | .60       |

# SUMMARY OF TEST RESULTS

## Recall Failures

### Reasons for Failure:

#### Indexing Language:

lack of term specificity  
term ambiguity  
false coordination  
failure to accommodate synonymous terms  
failure to accommodate synonymous phrases

#### Indexing:

term omission  
inappropriate term  
level of exhaustivity

#### Search Strategy:

inappropriate term  
too specific  
too general  
failure to cover all approaches

| MAI | KWOC |
|-----|------|
| 1   | 2    |
| 0   | 1    |
| 0   | 0    |
| 0   | 0    |
| 0   | 19   |
| 6   | 1    |
| 2   | 2    |
| 2   | 5    |
| 0   | 4    |
| 1   | 5    |
| 1   | 11   |
| 0   | 17   |

Exhibit B

# SUMMARY OF TEST RESULTS

## Relevance Failures

### Reasons for Failure:

#### Indexing Language:

lack of term specificity  
term ambiguity  
false coordination  
failure to accommodate synonymous terms  
failure to accommodate synonymous phrases

#### Indexing:

term omission  
inappropriate term  
level of exhaustivity

#### Search Strategy:

inappropriate term  
too specific  
too general  
failure to cover all approaches

| MAI | KWOC |
|-----|------|
| 18  | 12   |
| 14  | 7    |
| 2   | 0    |
| 0   | 0    |
| 0   | 19   |
| 1   | 1    |
| 1   | 1    |
| 4   | 0    |
| 0   | 17   |
| 0   | 0    |
| 23  | 21   |
| 0   | 1    |

Exhibit C

# SUMMARY OF TEST RESULTS

Number Of Recall And Relevance Failures Per Search

| Search # | MAI    |           | KWOC   |           |
|----------|--------|-----------|--------|-----------|
|          | Recall | Relevance | Recall | Relevance |
| 1        | 0      | 0         | 0      | 0         |
| 2        | 1      | 0         | 1      | 2         |
| 3        | 1      | 16        | 14     | 0         |
| 4        | 0      | 0         | 0      | 0         |
| 5        | 1      | 0         | 0      | 0         |
| 6        | 0      | 0         | 0      | 0         |
| 7        | 1      | 2         | 1      | 0         |
| 8        | 0      | 0         | 0      | 0         |
| 9        | 0      | 1         | 1      | 2         |
| 10       | 0      | 4         | 0      | 0         |
| 11       | 0      | 15        | 2      | 6         |
| 12       | 3      | 2         | 0      | 5         |
| 13       | 1      | 0         | 1      | 2         |
| 14       | 1      | 0         | 0      | 1         |
| 15       | 0      | 3         | 0      | 4         |
| 16       | 1      | 0         | 1      | 0         |
| 17       | 0      | 0         | 0      | 0         |
| 18       | 0      | 0         | 0      | 2         |
| 19       | 0      | 0         | 0      | 1         |
| 20       | 0      | 1         | 0      | 3         |

Exhibit D



## DISCUSSION

A considerable amount of work has been done in an attempt to determine the best methods for indexing a collection of documents. Best here is measured by the cost of indexing documents and maintaining any thesauri, recall and relevance figures, and the ease of the retrieval process. .

Lancaster investigated the comparative performance between searching index terms only or using index terms plus words in the title versus free text searching of abstract, title and index terms. Results show that recall increases 100% when manually assigned index terms are supplemented by free text searching of abstracts. Supplementation with free text searching of titles had little effect on recall.<sup>1</sup> Byrne investigated the relative merits of searching on titles, subject headings, abstracts, and free-language in the COMPENDEX data base. According to Byrne, the combination of terms from the titles and abstracts came closest to 100% retrieval, with searching of abstracts alone doing almost as well. Indexer input was found to be relatively unimportant.<sup>2</sup>

An indication of some difficulties encountered using free text searching of titles surfaces in Lefever's article.<sup>3</sup> Searching

<sup>1</sup>Frederick Lancaster, R.L. Rapport and J. Penry, "Evaluating the Effectiveness of an On-Line, Natural Language Retrieval System," Information Storage and Retrieval (October, 1972), p. 223-45.

<sup>2</sup>Jerry R. Byrne, "Relative Effectiveness of Titles, Abstracts, and Subject Headings for Machine Retrieval from the COMPENDEX Services," Journal of the American Society for Information Science (July-August, 1975), p. 223-29.

<sup>3</sup>Lefever, p. 339-42.

difficulties arising from the use of the uncontrolled indexing terminology produced by the KWOC are somewhat abated by the availability of a combined frequency count with scope notes and cross references.

The comparison of retrieval using MAI with a KWOC index is especially interesting in light of the fact that MAI provides vocabulary control, as well as indexing coverage of the narrative fields.

The KWOC index performed surprisingly well, equalling the recall and relevance capability of MAI. It seems odd that the KWOC index, taking terms only from the title and with no terminological control, did this well. Analysis of recall and relevance, relevance failures of the KWOC, however, does indicate that this KWOC is showing deficiencies pointed out in previous studies. Primarily, the lack of vocabulary control has a detrimental effect on retrieval.

The strong performance of the KWOC in this test suggests that KWOC technique may be useful to DTIC and KWOC indexing should be further tested in the Work Unit data base. I suggest that similar tests be done using a significantly larger sample of 1498s. In addition, the retrieval and recall performances of an index composed of terms assigned by MAI and a KWOC index of the title, abstract, or title and abstract, should be determined to indicate whether a combination of these two techniques will significantly improve retrieval.



# SELECTED BIBLIOGRAPHY

Cleverdon, Cyril; Mills, Jack; and Keen, Michael. ASLIB Cranfield Research Project: Factors Determining the Performance of Indexing Systems. Vol. 1: Test and Design. Cranfield Bedfordshire: College of Aeronautics, 1966.

Jahoda, Gerald. Information Storage and Retrieval Systems for Individual Researchers. New York: Wiley-Interscience, 1970.

Lancaster, Frederick W. Evaluation of the MEDLARS Demand Search Service. Bethesda, Md.: National Library of Medicine, 1968.

Full report of a two year evaluation of MEDLARS investigating adequacy of MEDLARS coverage of medical literature; quality of indexing; adequacy of indexing vocabulary; and ability of search strategies to respond to user needs. Provides MEDLARS recall and precision figures.

Lancaster, Frederick W. Information Retrieval Systems: Characteristics, Testing, and Evaluation. New York: John Wiley and Sons, Inc., 1968.

Stevens, Mary E. Automatic Indexing: A State-of-the-Art. National Bureau of Standards Monograph 91. Washington, D.C.: National Bureau of Standards, 1965.

Extensive review of KWOC and KWIC indexing.

Byrne, Jerry R. "Relative Effectiveness of Titles, Abstracts, and Subject Headings for Machine Retrieval from the COMPENDEX Services." Journal of the American Society for Information Science, (July-August, 1975), 223-29.

Compares the effectiveness of subject retrieval using titles, subject headings, abstracts, and free-language terms. Free text searching of abstracts found more effective than manually assigned subject headings.

Harris, Curt L., and Roberts, Suzanne H. "The Search for Tomorrow: Low Cost, Full Text Searching with Minimum Front End Investment." in American Society for Information Science. Proceedings of the ASIS Annual Meeting. Vol. 15: The Information Age in Perspective. White Plains, New York: Knowledge Industry Publications, Inc., 1978, 368-72.

Keen, E. Michael. "The Aberystwyth Index Languages Test." Journal of Documentation, 29:1 (March, 1973), 1-35.

Describes a comparison of the effectiveness and efficiency of five indexing languages in subject retrieval. No real differences observed.

Lancaster, Frederick W.; Rapport, R.L.; and Penry J. "Evaluating the Effectiveness of an On-line, Natural Language Retrieval System." Information Storage and Retrieval, 8:5 (October, 1972), 223-45.

Evaluation of Epilepsy Abstracts Retrieval System including a comparison of retrieval using free-text searching of title and abstract and retrieval using index terms.

Lefever, Maureen. "Managing an Uncontrolled Vocabulary Ex Post Facto." Journal of the American Society for Information Science, 23:6 (November-December, 1972), 339-42.

Describes vocabulary guides created to simplify searching KWOC indexes. Guides include frequency counts, cross references and scope notes.

Swanson, Don R. "Information Retrieval as a Trial-and Error Process." The Library Quarterly, 47:2 (April, 1977), 128-48.

Jacobs, Charles R. Machine-Aided Indexing: Technical Progress Report for Period July 1971-June 1972. DDC-TR-72-4 AD-754 400. Alexandria, Va.: Defense Documentation Center, 1972.

Klingbiel, Paul H. Machine-Aided Indexing: Technical  
Progress Report for Period Jan. 1967 - June 1969.  
DDC-TR-69-1 AD-696 200. Alexandria, Va.:  
Defense Documentation Center, 1969.